GAMES AND ECONOMIC BEHAVIOR 1, 191–209 (1989)

Two-Sided Matching with Incomplete Information about Others' Preferences

ALVIN E. ROTH

Department of Economics, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

A great deal of progress has been made recently in the study of two-sided matching processes, modelled both as cooperative games and as strategic games with complete information. Here we consider a natural model of incomplete information, in which agents know their own preferences but may not know those of others. It is shown that results concerning dominant and dominated strategies carry over from the complete information case to the present model, but results concerning Nash equilibria do not. Some of the modelling implications of this are briefly considered. © 1989 by Academic Press, Inc.

1. INTRODUCTION

There has been a great deal of progress in the study of two-sided matching markets modelled as games. Simple models without and with sidepayments were proposed by Gale and Shapley (1962) and Shapley and Shubik (1972), respectively, who studied the structure of the core. Recently more general models have been introduced in the literature,[1] and the strategic equilibria, as well as the core, have been examined.[2] There has also been progress in the use of these models to study the organization of particular labor markets (see, e.g., Roth (1984a, 1986) concerning the entry level market for American physicians).[3]

[1] See, e.g., Blair (1987), Crawford and Knoer (1981), Demange and Gale (1985), Demange *et al.* (1986), Kelso and Crawford (1982), Roth (1984c, 1985a), and Roth and Sotomayor (1988a).

[2] See, e.g., Dubins and Freedman (1981), Gale and Sotomayor (1985a,b), Graham and Marshall (1987), Leonard (1983), and Roth (1982, 1984b).

[3] Studies of several other labor markets are currently under way, the most closely related of which will be reported in Roth (1989). It seems likely that these models will prove most useful for studying markets in which relatively many heterogeneous workers and positions become available around the same time, e.g., entry level labor markets for elite professionals. But see also Mongell and Roth (1988).

0899-8256/89 $3.00

All of the above-mentioned papers formulate the problem at hand as a game of complete information. This involves at least the implicit assumption that the preferences of all agents are common knowledge. While this is clearly not an accurate description of the situation prevailing in the kinds of markets to which these models can mostly be applied, it may nevertheless be an appropriate way to model many of the features of these markets, since the alternative of modelling the market as a game of incomplete information presents other difficulties.

However, we will argue here that some of the conclusions reached about strategic decisions are particularly sensitive to the assumption of complete information, while others are not. Specifically, we will show that, for an appropriate model of incomplete information, the conclusions reached about dominant and dominated strategies continue to hold, while the conclusions about strategic equilibria in non-dominant strategies do not. In fact, it will be possible to show more than the simple negation of some of the complete information equilibrium results. It will be shown that things which happen at *all* equilibria of some complete information games of this sort need not happen at *any* equilibrium in games of incomplete information.

The organization of this paper is as follows. Section 2 describes the complete information model and the relevant results concerning dominant strategies and strategic equilibria for that case. Section 3 presents the incomplete information model and reconsiders equilibria and dominant strategies. The concluding section considers some open questions and briefly discusses some modelling issues.

2. The Complete Information Model

In order to concentrate on questions of information, we will consider the simplest of the two-sided matching models in the literature, the "marriage" market of Gale and Shapley (1962).

There are two finite, disjoint sets M and W: $M = \{m_1, m_2, \ldots, m_n\}$ is the set of men, and $W = \{w_1, w_2, \ldots, w_r\}$ is the set of women. Each man has preferences over the women, and each woman has preferences over the men. These preferences may be such that a man m would prefer to remain single rather than be married to some woman w, say, whom he does not care for.

To express these preferences concisely, the preferences of each man m will be represented by an ordered list of preferences, $P(m)$, on the set $W \cup \{m\}$. That is, a man m's preferences might be of the form

$$P(m) = w_2, w_1, m, w_3, \ldots, w_r,$$

indicating that his first choice is to be married to woman w_2, his second choice is to be married to woman w_1, and his third choice is to remain single. Similarly, each woman w in W has an ordered list of preferences, $P(w)$, on the set $M \cup \{w\}$. In examples an agent's preferences will be described by writing only the ordered set of people that the agent prefers to himself (herself), which are the acceptable matches for that agent. Thus the preferences $P(m)$ described above will be abbreviated by

$$P(m) = w_2, w_1.$$

An agent who is not indifferent about any two acceptable assignments (including remaining single) is said to have *strict* preferences. Let $\mathbf{P} = \{P(m_1), \ldots, P(m_n), P(w_1), \ldots, P(w_r)\}$ denote the set of preferences, one for each man and woman.

An outcome of the marriage market is a set of marriages. In general, not everyone may be married—some people may remain single. (We will adopt the convention that a person who is not married to someone is *self-matched.*) Formally we have

DEFINITION 1. A *matching* μ is a one-to-one correspondence from the set MUW onto itself of order two (that is, $\mu^2(x) = x$) such that if $\mu(m) \neq m$ then $\mu(m) \in W$ and if $\mu(w) \neq w$ then $\mu(w) \in M$. We refer to $\mu(x)$ as the *mate* of x.

Note that $\mu^2(x) = x$ means that if man m is matched to woman w (i.e., if $\mu(m) = w$), then woman w is matched to man m (i.e., $\mu(w) = m$). The definition also requires that individuals who are not single be matched with agents of the opposite set—i.e., men are matched with women. These two requirements explain why matchings can be thought of as sets of "marriages."

A matching will sometimes be represented as a set of a matched pairs. Thus, for example, the matching

$$\mu: \begin{array}{ccccc} w_4 & w_1 & w_2 & w_3 & (m_5) \\ m_1 & m_2 & m_3 & m_4 & m_5 \end{array}$$

has m_1 married to w_4 and m_5 remaining single, i.e., $\mu(m_1) = w_4$ and $\mu(m_5) = m_5$.

Each agent's preferences over alternative matchings correspond exactly to his (her) preferences over his own mates at the two matchings. Thus man m, say, prefers matching μ to matching ν if and only if he prefers $\mu(m)$ to $\nu(m)$. Thus we are assuming that man m cares about who *he* is matched with, but is not otherwise concerned with the mates of other agents. The rules of the game are that any man and woman may marry one another if they both agree, and any individual is free to remain single.

Consider a matching μ which matches a pair (m, w) who are not mutually acceptable. Then at least one of the individuals m and w would prefer to be single rather than be matched to the other. Such an individually irrational matching μ will be said to be *blocked* by the unhappy individual. Consider a matching μ such that there exist a man m and a woman w who are not matched to one another at μ, but who prefer each other to their assignments at μ. The man and woman (m, w) will be said to *block* the matching μ.

DEFINITION 2. A matching μ is *stable* if it is not blocked by any individual or any pair of agents.

It is easy to show that, for this model, the set $S(\mathbf{P})$ of stable matchings coincides with the core.[4]

For a given marriage market, we say that a stable matching μ is an *M-optimal stable matching* if every man likes it at least as well as any other stable matching. We can define a W-optimal stable matching similarly. Gale and Shapley (1962) proved the following.

THEOREM 1. *A stable matching exists for every marriage market. Furthermore, when all men and women have strict preferences, there always exist an M-optimal stable matching and a W-optimal stable matching.*

Of course when preferences are strict the M- and W-optimal stable matchings are unique and will be denoted by μ_M and μ_W, respectively.

2.1. *Strategic Questions*

Consider a marriage market $(M, W, \mathbf{P})$ whose outcome will be determined by some "revelation mechanism" based on a list of preference orderings which agents will state. That is, each man m, whose preferences are $P(m)$, is faced with the strategy choice of what preference ordering $Q(m)$ to state, and each woman w with preferences $P(w)$ must state a preference ordering $Q(w)$. The set of stated preference lists, one for each man and woman, will be denoted by $\mathbf{Q} = \{Q(m_1), \ldots, Q(m_n), Q(w_1), \ldots, Q(w_r)\}$. The mechanism produces a matching $\mu = h(\mathbf{Q})$, where h is the function that describes the mechanism's output for any set $\mathbf{Q}$ of stated preferences. The mechanism h together with the data $(M, W, \mathbf{P})$ constitutes a strategic game, in which the strategies of the agents are

[4] For the simplest model of many-to-one matching, the set of stable matchings is the subset of the core corresponding to those matchings that are not weakly dominated (Roth, 1985b).

what preferences to state, and we will sometimes say that this is the game "induced" by h.

For simplicity in what follows, we will consider the case in which all preferences are strict.[5] This means not only that the true preferences $\mathbf{P}$ of the agents include no indifferences, but also that only strict preferences may be stated to the revelation mechanism. A mechanism h that for any stated preferences $\mathbf{Q}$ produces a matching $h(\mathbf{Q})$ that is stable with respect to $\mathbf{Q}$ is called a *stable mechanism.* If $h(\mathbf{Q})$ equals the M-optimal stable matching with respect to $\mathbf{Q}$, then h is the *M-optimal* stable mechanism.

We now consider the strategic results for the complete information case that will be reexamined in the case of incomplete information.[6]

THEOREM 2. Impossibility Theorem. *When there are at least two agents on each side of the market, no stable matching mechanism exists which always makes stating the true preferences a dominant strategy for every agent.*

Thus any stable matching mechanism must sometimes present at least some agents with non-trivial strategic decisions. The next theorem shows that, at least in the case of one-to-one matching studied here, it is possible to organize the market so that only the agents on one side face these strategic difficulties, by employing the mechanism which yields the optimal stable matching for the agents on the other side of the market.[7]

THEOREM 3. *The M-optimal stable mechanism makes it a dominant strategy for each man to state his true preferences. (Similarly, the W-optimal stable mechanism makes it a dominant strategy for every woman to state her true preferences.)*

So when an M-optimal stable mechanism is employed, Theorems 2 and 3 together imply that it is the women who will sometimes have difficult strategic decisions to make. The following theorem shows that at least the decision of who to list first on their stated preference lists is not a difficult one. The theorem speaks of a woman w's "true first choice," meaning the individual who is first on her true preference list $P(w)$ (recall that preferences are assumed to be strict).

THEOREM 4. *When the M-optimal stable mechanism is used, any strategy $Q(w)$ in which w does not list her true first choice at the head of her list is dominated by the strategy $Q'(w)$ which does list her true first choice first, but otherwise leaves $Q(w)$ unchanged.*

[5] But see Roth (1984b) for a discussion of the non-strict case.

[6] Theorems 2–4 are from Roth (1982), and Theorem 3 was independently presented by Dubins and Freedman (1981).

[7] This result does not apply to agents who may be matched to more than one agent simultaneously, in models of many-to-one matching (Roth, 1985a; Sotomayor, 1987).

Turning from dominant strategies to equilibrium strategies, we have the following results for the complete information case.[8]

THEOREM 5. *Suppose each man chooses his dominant strategy and states his true preferences, and the women choose any set of strategies (preference lists) $P'(w)$ that form an equilibrium for the matching game induced by the M-optimal stable mechanism. Then the corresponding M-optimal stable matching for* $(M, W, \mathbf{P}')$ *is one of the stable matchings of* $(M, W, \mathbf{P})$.

THEOREM 6. *When all preferences are strict, let μ be any stable matching for* $(M, W, \mathbf{P})$. *Suppose each woman w in $\mu(M)$ chooses the strategy of listing only $\mu(w)$ on her stated preference list of acceptable men (and each man states his true preferences). This is an equilibrium in the game induced by the M-optimal stable matching mechanism (and μ is the matching that results).*

THEOREM 7. *Let* $\mathbf{P}'$ *be a set of preferences in which each man states his true preferences, and each woman states a preference list which ranks the men in the same order as her true preferences, but ranks as unacceptable all men who are ranked below $\mu_W(w)$. These preferences* $\mathbf{P}'$ *are a strong equilibrium for the women*[9] *in the game induced by the M-optimal stable matching mechanism (and μ_W is the matching that results).*

So, while Theorem 2 asserts that at equilibrium some agents can be expected to state preferences different from their true preferences, Theorem 5 (together with Theorem 3) shows that at any equilibrium in undominated strategies, the matching that results will nevertheless be stable with respect to the true preferences, when the M-optimal stable mechanism is used. Theorems 6 and 7 give examples of two such equilibria. Note that to implement either of these equilibria the women must have a good deal of information about other agents' preferences, since this is needed in order to determine whether a particular matching is in the core. It is this observation which most directly motivates the present investigation, since in the markets to which this kind of theory has mostly been applied, the agents do not possess detailed information about the preferences of others, although they may possess some general information about these preferences.

A related result, which implies the conclusions of Theorem 3 in the complete information case, is the following.[10]

[8] Theorem 5 is from Roth (1984b), and Theorems 6 and 7 are from Gale and Sotomayor (1985b).

[9] That is, an equilibrium with the property that no coalition of women can do better by changing their strategies.

[10] Theorem 8 was first proved by Dubins and Freedman (1981). A short proof is given by Gale and Sotomayor (1985a).

THEOREM 8. *Let* $\mathbf{P}$ *be the true preferences of the agents, and let* $\overline{\mathbf{P}}$ *differ from* $\mathbf{P}$ *in that some coalition* $\overline{M}$ *of the men falsify their preferences. Then there is no matching* μ, *stable for* $\overline{\mathbf{P}}$, *which is preferred to* μ_M *by all members of* $\overline{M}$.

A result that will be useful in the proofs is the following:[11]

THEOREM 9. *In a market* $(M, W, \mathbf{P})$ *with strict preferences, the set of people who are single is the same for all stable matchings.*

3. Incomplete Information about Others' Preferences

We now consider a model in which each agent does not know the preferences of the others, but knows only the probability distribution from which these are drawn. Since we will be dealing with probabilities, we will need to consider not merely the preference orderings of the players, but also their expected utility functions. (Preferences will still be assumed to be strict.) We will also consider games with quite general rules and strategy sets, and not merely revelation games in which agents simply state their preferences. A general matching game with incomplete information about others' preferences will be given by a collection.

$$\Gamma = (N = M \cup W, \{D_i\}_{i \in N}, g, U = X_{i \in N} U_i, F).$$

The set N of players consists of the men and women to be matched. The sets D_i describe the decisions facing each player in the course of any play of the game (i.e., an element d_i of D_i specifies the action of player i at each point in the extensive form of the game at which he has decisions to make). The function g describes how the actions taken by all the agents correspond to matchings and lotteries over matchings, i.e., g: $X_{i \in N} D_i \rightarrow L[\mathcal{M}]$, where $\mathcal{M}$ is the set of all matchings between the sets M and W, and $L[\mathcal{M}]$ is the set of all probability distributions (lotteries) over $\mathcal{M}$. The set U_i is the set of all utility functions defined over the possible mates for player i and the possibility of remaining single, and F is a probability distribution over n-tuples of utility functions $u = \{u_i\}_{i \in N}$, for u_i in U_i. The interpretation is that a player's "type" is given by his utility function, and at the time in which the players must choose their strategies each player knows his own type, and the probability distribution F over vectors u is common knowledge. The special case of a game of complete information

[11] The first statement of Theorem 9 of which I am now aware appears in McVitie and Wilson (1970) for the case in which all men and women are mutually acceptable. It was also proved in Roth (1984a) in a context directly applicable to the more general case considered here, and with a short proof by Gale and Sotomayor (1985a).

occurs when the distribution F gives a probability of one to some vector u of utilities. While the set U of possible utilities for the players is large, we will typically be concerned with games in which only a subset of U has positive probability. For simplicity, we will henceforth confine our attention to cases in which the set of utility functions which occur with positive probability is countable. In any event, since each player i knows his own utility function u_i, he can compute a conditional probability $p_i(u_{-i}|u_i)$ for each vector of other players'. utilities u_{-i} in $U_{-i} \equiv X_{j \neq i} U_i$, by applying Bayes' rule to F.

This is not the most general kind of incomplete information model we might consider (cf. Harsanyi, 1967, 1968a,b). The only unknown information is the other players' utilities. In particular, players know their own utilities for being matched with one another even though they do not know what "type" the other is: the final matching, and hence each player's utility payoff, depends only on the actions of the players, not on their types. (That is, each player's utility payoff depends on his own type, and on the actions of all the players, but not on the types of the other players.) Stated another way, players' types do not affect their desirability, only their desires. This seems like a natural assumption for elite professional markets for entry level positions. For example, in the hospital intern market, after the usual interviewing has been completed, top students are able to rank prestigious programs, and vice versa. But agents do not know, for example, how their top choices rank *them.*[12]

A *strategy* for player i is a function σ_i from his type (which in this case is his utility function) to his decisions, i.e., σ_i: $U_i \to D_i$. If $\sigma = \{\sigma_i\}_{i \in N}$ denotes the strategy chosen by each player, then for each vector u of players' utility functions, $\sigma(u) = \{d_i \in D_i\}_{i \in N}$ describes the decisions made by the players, which result in the matching (or lottery over matchings) $g(\sigma(u))$. Consequently a set of strategy choices σ results in a lottery over matchings, whose probabilities are determined by the probability distribution F over vectors u, and by the function g. The expected utility to player i who is of type u_i is given by

$$u_i(\sigma) = \Sigma_{u_{-i} \in U_{-i}} p_i(u_{-i}|u_i) u_i[g(\sigma(u_{-i}, u_i))].$$

An *equilibrium* set of strategies is a σ^* such that, for all players i in N and all utility functions u_i in U_i, $u_i(\sigma^*) \geq u_i(\sigma^*_{-i}, \sigma_i)$ for all other strategies σ_i for player i. That is, when player i's utility is u_i the strategy σ_i^* determines player i's decision $d_i^* = \sigma_i^*(u_i)$, and the equilibrium condition requires that for all players i and all types u_i which occur with positive

[12] Note the difference between this kind of model and those in the job-search literature, in which the interviewing process itself is modelled, so that in effect agents are uncertain about their own preferences.

probability, player i cannot profitably substitute another decision $d_i = \sigma_i(u_i)$.

3.1. *Revelation Games*

We now consider the class of incomplete information games called "revelation games," in which players are required only to state ("reveal") their types, which in this case are their utilities.[13] Recall that a general matching game with incomplete information about others' preferences is given by $\Gamma = (N = M \cup W, \{D_i\}_{i \in N}, g, U = X_{i \in N} U_i, F)$. We may call $[\{D_i\}_{i \in N}, g]$ the *mechanism*, and $[U, F]$ the *state of information* of the game. Then a game Γ is specified by a set of players, a mechanism, and a state of information. A *revelation game* Γ_{R} is a game in which the mechanism is of the form $[\{D_i = U_i\}_{i \in N}, h]$, where h is a function which takes stated preferences into matchings or lotteries over matchings, i.e., h: $X_{i \in N} U_i \rightarrow L[\mathcal{M}]$. We will sometimes call the function h itself a *revelation mechanism,* it being understood that the decision facing each agent is simply what utility to state. We will pay particular attention to the strategy of truth telling in a revelation game, i.e., the strategy $\sigma_i^{\mathrm{T}}(u_i) = u_i$ in which each agent states his true type.

For any general game Γ of incomplete information, and any equilibrium σ^* of Γ, we can define the *revelation game corresponding to* σ^* to be the game $\Gamma_{\mathrm{R}}(\sigma^*)$ with the same set of players and state of information as Γ and with the revelation mechanism h given by $h(u) = g(\sigma^*(u))$. That is, the revelation mechanism h takes any set of stated utilities and produces the same matching (or lottery over matchings) as that which would have been produced by the equilibrium σ^* in the game Γ if the true utilities of the players had been u. (That is, regardless of the actual types of the players in the corresponding revelation game, if they collectively state the utilities u the resulting matching is the one that would have resulted in Γ under strategies σ^* if the vector u corresponded to the true player types.) The following observation, which is widely used in proofs about games of incomplete information, is known as the "revelation principle" (see, e.g., Myerson, 1985).

The revelation principle. For any equilibrium σ^* of a general incomplete information matching game Γ, let $\Gamma_{\mathrm{R}}(\sigma^*)$ be the corresponding revelation game. Then

1. Truth telling is an equilibrium. That is the strategies $\sigma^{\mathrm{T}} = \{\sigma_i^{\mathrm{T}}\}_{i \in N}$ are an equilibrium in $\Gamma_{\mathrm{R}}(\sigma^*)$.

[13] We will later also consider the special case of revelation games in which players can only state preference orderings, and not numerical utilities. This appears to be the situation in most of the actual two-sided matching markets which employ revelation procedures.

2. When all players tell the truth in $\Gamma_R(\sigma^*)$, the resulting matching (or lottery over matchings) is the same as that when the players play the strategies σ^* in Γ.

The second observation follows immediately from the definition of the revelation mechanism h and does not depend on whether or not the strategies σ^* are an equilibrium in the original game Γ. The first observation, that truth telling is an equilibrium in the revelation game, follows from the fact that σ^* is an equilibrium in Γ: if player i with utility u_i could profit from stating another utility function v_i when all other agents state the truth in the revelation game, then he could get the same outcome in game Γ, and hence also profit, from playing $\sigma(v_i) \equiv d_i$ instead of $\sigma^*(u_i) \equiv d_i^*$ when all other players use the strategies σ^*. But then σ^* would not be an equilibrium, contrary to assumption.

We will call a revelation mechanism h *stable* if for any stated utilities u its output $h(u)$ is a stable matching or lottery over stable matchings, i.e., if $h(u) \in L[S(u)]$. Note that the set of stable matchings is sensitive only to the ordinal preferences; i.e., if $u = (u_1, \ldots, u_{n+r})$ is a vector of expected utility functions, one for each agent, then there is a unique vector $\mathbf{P} = \mathbf{P}(u)$ of ordinal preferences corresponding to these utilities, and the set of stable matchings is the same for any two utility vectors that have the same corresponding preferences (i.e., $S(u) = S(v) = S(\mathbf{P})$ whenever $\mathbf{P}(u) = \mathbf{P}(v) = \mathbf{P}$).

Of course while the set of stable matchings responds only to the ordinal information contained in the expected utility functions, a mechanism for selecting a stable matching can depend on the expected utilities in a more detailed way. However, in the complete information case we considered mechanisms which respond only to stated preferences. This corresponds to what is generally observed in markets which employ centralized matching mechanisms, such as the market for American medical interns (cf. Roth, 1984a), the similar regional markets for pre-registration positions in the United Kingdom (cf. Roth, 1989), or the bid matching procedure used by American sororities (cf. Mongell, 1988; Mongell and Roth, 1988). The M-optimal and W-optimal stable mechanisms, for example, depend only on the preferences, not on the utilities. Formally, with respect to a set MUW of agents we can define the following special class of revelation mechanisms.

DEFINITION 3. An *ordinal* stable matching mechanism is a function defined on all utility vectors u such that $h(u)$ is in $S(\mathbf{P})$ where $\mathbf{P}$ are the preferences corresponding to u, and such that if v is another utility vector corresponding to $\mathbf{P}$, then $h(u) = h(v)$.

Note that when an ordinal mechanism is used, it is equivalent to thinking of agents stating either their utility functions or the corresponding preferences. In contrast to ordinal stable mechanisms, the general stable revelation mechanisms we have defined could be called *random cardinal* stable mechanisms, in recognition of the fact that they may yield different (and random) outcomes for stated utilities u and v for which $\mathbf{P}(u) = \mathbf{P}(v)$.

3.2. *Equilibrium and Stability*

The first result is an impossibility theorem that provides a strong negation to the conclusions of Theorem 5 about equilibria in the complete information case when the M-optimal stable mechanism is employed. It says that, in the incomplete information case, no equilibrium of any mechanism can have the stability properties that every equilibrium of the M-optimal stable mechanism has in the complete information case. The strategy of the proof will be to observe that, by the revelation principle, if any such mechanism existed then there would be a stable revelation mechanism with truth telling as an equilibrium, and then to show that no such revelation mechanism exists. The proof of Theorem 10 thus also shows that the impossibility result of Theorem 2 generalizes to the case of random cardinal mechanisms and incomplete information.

THEOREM 10. *If there are at least two agents on each side of the market, then for any general mechanism* $[\{D_i\}_{i\in N}, g]$ *there exist states of information* $[U, F]$ *for which every equilibrium* σ *of the resulting game* Γ *has the property that* $g(\sigma(u)) \notin L[S(u)]$ *for some* $u \in U$. *(And the set of such* u *with* $g(\sigma(u)) \notin L[S(u)]$ *has positive probability under* F.*) That is, there exists no mechanism with the property that at least one of its equilibria is always stable with respect to the true preferences at every realization of a game.*

To see that at least two agents are required on each side of the market, note that, since preferences are strict, a game with only one agent on one side of the market must have a unique stable matching, at which the singleton agent gets his or her highest ranked mutually acceptable choice. In such a game it is not hard to see that it is a dominant strategy for all agents to state their true preferences when a stable mechanism is used.

The proof will examine the smallest remaining case, of two agents on each side of the market, and give an example, i.e., a state of information, that causes every stable revelation mechanism (and hence every mechanism) to fail. The conclusion for larger sets of agents follows from the fact that the four agents who play a role in the proof can be embedded in any larger set of agents without affecting the conclusion, so long as their preferences are not changed (and so that, in particular, in a larger market these four agents do not consider any additional agents to be acceptable matches).

The example. In order not to obscure the basic simplicity of the proof, it will be helpful to state the ordinal preferences of the agents first, and consider their utility functions later. The agents are $N = M \cup W$, where $M = \{m_1, m_2\}$ and $W = \{w_1, w_2\}$. The most likely distribution of player types corresponds to the following preferences:

$$P(m_1) = w_1, w_2, \qquad P(w_1) = m_2, m_1$$
$$P(m_2) = w_2, w_1, \qquad P(w_2) = m_1, m_2.$$

Agents m_2 and w_2 have no other types which occur with positive probability, but m_1 and w_1 may each have two types, with their other possible preferences being

$$P'(m_1) = w_1, \qquad P'(w_1) = m_2.$$

The probability that m_1 has preferences $P'(m_1)$ rather than $P(m_1)$ is q, which is also the probability that w_1 has preferences $P'(w_1)$ rather than $P(w_1)$.

Let

$$\begin{aligned} \mathbf{P} &= (P(m_1), P(m_2); P(w_1), P(w_2)),\\ \mathbf{P}' &= (P'(m_1), P(m_2); P(w_1), P(w_2)),\\ \mathbf{P}'' &= (P(m_1), P(m_2); P'(w_1), P(w_2)), \text{ and}\\ \mathbf{P}''' &= (P'(m_1), P(m_2); P'(w_1), P(w_2)) \end{aligned}$$

be the various preference profiles which can arise with probabilities $(1 - q)^2$, $q(1 - q)$, $q(1 - q)$, and q^2, respectively.

We will suppose that, for each type of each agent, the utility of being matched to his first choice is 2, to his second choice is 1, and to his third choice is 0. Thus, for example, when m_1 has preferences $P(m_1)$ his utility is $u_{m_1}(w_1) = 2$, $u_{m_1}(w_2) = 1$, and $u_{m_1}(m_1) = 0$, and when he is the type with preferences $P'(m_1)$ his utility is given by $u'_{m_1}(w_1) = 2$, $u'_{m_1}(m_1) = 1$, and $u'_{m_1}(w_2) = 0$.

There are three distinct matchings that may be stable for some realization of the possible types which are denoted by

$$\mu = \begin{pmatrix} m_1 & m_2 \\ w_1 & w_2 \end{pmatrix}, \qquad \nu = \begin{pmatrix} m_1 & m_2 \\ w_2 & w_1 \end{pmatrix}, \qquad \tau = \begin{pmatrix} m_1 & m_2 & (w_1) \\ (m_1) & w_2 & w_1 \end{pmatrix}.$$

The sets of stable matchings corresponding to each of the possible combinations of types are

$$S(\mathbf{P}) = \{\mu, \nu\}, \qquad S(\mathbf{P}') = \{\mu\},$$
$$S(\mathbf{P}'') = \{\nu\}, \qquad S(\mathbf{P}''') = \{\tau\}.$$

Proof of Theorem 10. Note that this example is symmetric between the two sides of the market. Since $S(\mathbf{P})$ contains only two matchings, we may suppose without loss of generality that an arbitrary stable revelation mechanism h chooses the stable matching μ with probability at least one-half when the utilities u corresponding to the preferences $\mathbf{P}$ are stated. (Recall that there are only two stable matchings in this case, so a stable mechanism must choose one of them with a probability of at least one-half: If h instead chose ν with probability greater than one-half, the argument which follows would proceed with m_2 replacing w_2.) We will show that h gives w_2 an incentive to mis-state her utility when the other agents adopt the strategy of truth telling and that therefore truth telling is not an equilibrium for any stable revelation mechanism when the state of information is as given in the example.

Let σ be a strategy 4-tuple, and denote by $\sigma|\mathbf{P}$ the vector of stated utilities that results when the true types of the players correspond to the preferences $\mathbf{P}$, and similarly denote by $\sigma|\mathbf{P}'$, $\sigma|\mathbf{P}''$, and $\sigma|\mathbf{P}'''$ the stated utilities that correspond to the other configurations of player types. Then the expected utility of w_2, whose utility function is u_{w_2} as specified earlier, is

$$u_{w_2}(\sigma) = (1 - q)^2 u_{w_2}(h(\sigma|\mathbf{P})) + q(1 - q)[u_{w_2}(h(\sigma|\mathbf{P}')) \\ + u_{w_2}(h(\sigma|\mathbf{P}''))] + q^2 u_{w_2}(h(\sigma|\mathbf{P}''')).$$

So, if q is sufficiently small, any potential losses in states of the world other than $\mathbf{P}$ are offset in expected utility by any gains in the most probable state of the world, $\mathbf{P}$. Thus for $q = 0.01$, say, w_2 prefers to get ν for certain instead of μ with a probability of at least one-half in state $\mathbf{P}$, at the cost of getting any other match in any of the other states. But w_2 can achieve this when all other agents adopt the truth-telling strategy, by stating any utility function corresponding to the preferences $P'(w_2) = m_1$, which are not her true preferences. The reason is that when all other agents state utilities corresponding to $\mathbf{P}$, but w_2 states a utility corresponding to $P'(w_2)$, then the unique stable matching is ν, and so any stable mechanism must choose ν in this case. So truth telling is not an equilibrium, since when other agents adopt truth-telling strategies, w_2 will prefer to state preferences corresponding to $P'(w_2)$. (Since w_2 has only one type with positive probability, this fully describes her strategy, as far as it affects any agent's expected utility.)

By the revelation principle, the proof is now complete, since if any mechanism existed with an equilibrium that always produced a stable outcome, the corresponding revelation mechanism would be a stable mechanism in which truth telling was an equilibrium. ∎

The next theorem shows that the conclusion of Theorem 8 also does not generalize to the case of incomplete information. It is possible for coalitions of men, by mis-stating their preferences, to obtain a preferable matching (even) from the M-optimal stable mechanism. This is so even though, as we will see in Theorem 12, it remains a dominant strategy for each man to state his true preferences.

THEOREM 11. *In games of incomplete information about preferences, the M-optimal stable mechanism may be group manipulable by the men.*

Proof. Consider the following example:[14] The agents are $N = M \cup W$, where $M = \{m_1, m_2, m_3\}$ and $w = \{w_1, w_2, w_3, w_4\}$. The most likely distribution of player types corresponds to the following preferences:

$$\begin{array}{ll} P(m_1) = w_4, w_2, w_1, w_3, & P(w_1) = m_1, m_2, m_3 \\ P(m_2) = w_4, w_1, w_2, w_3, & P(w_2) = m_3, m_1, m_2 \\ P(m_3) = w_1, w_2, w_3, & P(w_3) = m_1, m_2, m_3 \\ & P(w_4) = w_4. \end{array}$$

Note that, under these preferences, woman w_4 is unwilling to be matched with any man. Except for w_4, each agent has only one type having positive probability, corresponding to the above preferences. Woman w_4 has two types having positive probability, with her other possible preferences being $P'(w_4) = m_1, m_2$.

The (small) probability that w_4 has preferences $P'(w_4)$ rather than $P(w_4)$ is q. Let $\mathbf{P} = (P(m_1), P(m_2), P(m_3), P(w_1), \ldots, P(w_4))$ and $\mathbf{P}' = (P(m_1), \ldots, P'(w_4))$ denote the two preference profiles which can arise, with probabilities $(1 - q)$ and q, respectively.

Let $\mu_M(\mathbf{P})$ and $\mu_M(\mathbf{P'})$ denote, respectively, the M-optimal stable matchings with respect to $\mathbf{P}$ and $\mathbf{P}'$. Then

$$\mu_M(P) = \begin{pmatrix} m_1 & m_2 & m_3 & (w_4) \\ w_1 & w_3 & w_2 & w_4 \end{pmatrix} \quad \text{and}$$

$$\mu_M(P') = \begin{pmatrix} m_1 & m_2 & m_3 & (w_3) \\ w_4 & w_1 & w_2 & w_3 \end{pmatrix}.$$

For q sufficiently small,[15] the coalition of m_1 and m_2 can assure itself a higher expected utility by stating the preferences $Q(m_1) = w_2, w_1, w_3$ and $Q(m_2) = w_4, w_3$ when other agents all state their true preferences. This is because

[14] Which is adapted from Roth (1982) by the addition of agent w_4.

[15] Specifically, for $q < [u_{m_1}(w_2) - u_{m_1}(w_1)]/[u_{m_1}(w_4) - u_{m_1}(w_1)]$.

$$\mu_M(\mathbf{Q}|\mathbf{P}) = \begin{pmatrix} m_1 & m_2 & m_3 & (w_4) \\ w_2 & w_3 & w_1 & w_4 \end{pmatrix} \quad \text{and}$$

$$\mu_M(\mathbf{Q}|\mathbf{P}') = \begin{pmatrix} m_1 & m_2 & m_3 & (w_3) \\ w_2 & w_4 & w_1 & w_3 \end{pmatrix}.$$

Note that both men m_1 and m_3 profit from m_2's misrepresentation when the true preferences are **P** and that m_2 gets the same spouse as if he had stated his true preferences in this case.[16] But m_2 does better when the preferences are **P′** than he would have if m_1 had stated his true preferences, and although m_1 does worse in this case, he nevertheless receives a higher expected utility when he and m_2 both mis-state their preferences according to **Q**, since the probability q is small. Note that when they must state their preferences, neither m_1 nor m_2 knows whether the true preferences are **P** or **P′**.

The fact that, even in the case of complete information, it is possible for a coalition of men to mis-state their preferences in a way that does not hurt any of them and helps some of them means that the conclusion from Theorem 8 that coalitions of men cannot collectively manipulate the M-optimal mechanism to their advantage cannot be expected to be very robust. Once there is any possibility that the men can make any sort of sidepayments among themselves, this conclusion is no longer justified. Theorem 11 shows that uncertainty about the preferences of other agents allows some transfers in an expected utility sense, with men able to trade a gain in one realization for a gain in another. Note that this is so for any positive q, i.e., even when q is arbitrarily small, in which case there is very little uncertainty about the preferences.

3.3. *Dominant and Dominated Strategies*

In contrast to the results for equilibria, the results concerning dominant strategies in the complete information case do generalize to the case of incomplete information. We begin with a general proposition about the relationship of dominant strategies for complete information revelation games and the corresponding revelation games of incomplete information about others' preferences. (Here "corresponding" means having the same mechanism and set of players, but allowing for different states of information.)

PROPOSITION 1. *If h is a mechanism that makes stating his true utility u_i a dominant strategy for player i in every complete information revelation game Γ (i.e., for every specification of u), then the truth-telling strat-*

[16] In Roth (1982) the point was that in the complete information case with preferences **P**, man m_2 could help the other men at no cost to himself.

egy $\sigma_i^T(u_i) = u_i$ is a dominant strategy in any corresponding game Γ^* of incomplete information about others' preferences.

Proof. Suppose not. Then there is another strategy σ_i which does better for at least some realization $u \in U$, such that $\sigma_i(u_i) \neq u_i$. But this contradicts the fact that, in the complete information game with the utilities given by u, it is a dominant strategy for agent i to state u_i. ■

The following is an immediate consequence of Proposition 1 and Theorem 3.

THEOREM 12. *In matching with incomplete information about others' preferences, the M-optimal stable mechanism makes it a dominant strategy for each man to state his true preferences; i.e., $\sigma_i^T(u_i) = u_i$ is a dominant strategy for each man. (Similarly, the W-optimal stable mechanism makes it a dominant strategy for every woman to state her true preferences.)*

A similar, pointwise argument on realizations of the types of players allows us to prove the following parallel to Theorem 4.

THEOREM 13. *When an M-optimal stable mechanism is used in matching with incomplete information about others' preferences, any strategy $\sigma_i(u_i)$ for a woman w_i is dominated if her stated first choice is not her true first for each u_i in U_i.*

Proof. Consider a strategy σ_i for some woman w_i that for at least one u_i in U_i states a utility $\sigma_i(u_i)$ which ranks highest some alternative different from the highest ranked alternative according to u_i, i.e., such that the maximum of $\sigma_i(u_i)$ over $M \cup \{w_i\}$ is achieved by some alternative other than the one which maximizes u_i, which we will denote s^*. Let $\sigma_i^*: U_i \rightarrow D_i$ be a strategy that differs from σ_i only for $u_i \in U_i$. Furthermore, suppose that the numbers $\sigma_i^*(u_i)(s) = \sigma_i(u_i)(s)$ for all $s \neq s^*$ in $M \cup \{w_i\}$, and that s^* maximizes $\sigma_i^*(u_i)$. Recall that the matching that results, and the corresponding utility that w_i derives, depends only on the stated utilities of the players (their actions) and not on their types. So by Theorem 4 woman w_i does at least as well by playing σ_i^* as she does by playing σ, for any strategy choices of the other agents, and strictly better for at least one set of other agents' strategy choices. ■

Since the M-optimal stable mechanism is an ordinal mechanism, no difficulty in the proofs arises from the fact that Theorem 3 and 4 were stated in terms of a game in which agents state preferences, while Theorems 12 and 13 are stated in terms of a game in which agents state utilities.

Proposition 1 and the related argument in the proof of Theorem 13 illustrate a kind of "dominant strategy principle," connecting dominant strategy results for classes of complete information games to parallel results for incomplete information games, as exemplified by Theorems 12 and 13. As we have seen, no such parallels exist for the equilibrium results for the complete information game.

4. Discussion

Since the questions in this paper are motivated to a large extent by empirical questions concerning the behavior of agents in real markets, it seems appropriate to conclude with some comments on modelling issues. The first of these concerns the definition of stability, which was carried over unchanged from the complete information model to the incomplete information model. In the context of the incomplete information model, the kind of stability studied here is ex post stability, in the sense that a stable matching would remain stable even if all the preferences were to become common knowledge.

The reason this is not an excessively strong requirement is that, when a matching is proposed, each agent knows which alternative he prefers. Although he does not yet know how his preferred alternatives evaluate him, the ease of ascertaining this in a number of the markets of interest is precisely what causes the instability observed in those markets. For example, in the hospital intern labor market, a medical student who has been offered an internship at his third choice hospital can easily contact his first and second choices to see if they prefer him to any of the students they are currently considering. In the late 1940s, prior to the introduction of a stable matching procedure in this market, just this kind of very late search caused verbal and other contracts to be broken in sufficiently substantial numbers to interfere with the operation of the market (see Roth, 1984a).

A larger modelling issue is the question of when complete and incomplete information models are most useful. It seems clear that, in most of the markets to which these models can be applied, agents do not know with precision the preferences of all the other agents. However, it is rarely apparent what, if any, priors about agents' preferences can be reasonably described as being shared by all agents. So both kinds of models impose costs. For certain kinds of questions about stability, its ex post nature makes the differences between the two kinds of models unimportant. For other questions, there seems to be no practical alternative to examining both kinds of models and the answers they give, in the light of how much they seem to make strained assumptions about the markets in question. In just this way, the information required to implement the equilibria identified by the complete information model led to the present exploration of the incomplete information case. However, the equilibrium results obtained here are negative, and it remains an open question what general positive characterizations of equilibria can be obtained in the incomplete information case, or which of the many properties of this and related complete information models will generalize to the case of incomplete information. (For a comprehensive survey which concentrates on the complete information case, see Roth and Sotomayor (1988b).)

Acknowledgments

It is a pleasure to acknowledge helpful conversations on this topic with Marilda Sotomayor and Shmuel Zamir. This work has been partially supported by grants from the National Science Foundation, the Office of Naval Research, and the Alfred P. Sloan Foundation.

References

Blair, C. (1987). "The Lattice Structure of the Set of Stable Matchings with Multiple Partners," *Math. Oper. Res.*, in press.

Crawford, V. P., and Knoer, E. M. (1981), "Job Matching with Heterogeneous Firms and Workers," *Econometrica* **49,** 437–450.

Demange, G., and Gale, D. (1985). "The Strategy Structure of Two-Sided Matching Markets," *Econometrica* **53,** 873–888.

Demange, G., Gale, D., and Sotomayor, M. (1986). "Multi-item Auctions," *J. Polit. Econ.* **94,** 863–872.

Dubins, L. E., and Freedman, D. A. (1981) "Machiavelli and the Gale–Shapley Algorithm," *Amer. Math. Monthly* **88,** 485–494.

Gale, D., and Shapley, L. (1962). "College Admissions and the Stability of Marriage," *Amer. Math. Monthly* **69,** 9–15.

Gale, D., and Sotomayor, M. (1985a). "Some Remarks on the Stable Matching Problem," *Discrete Appl. Math.* **11,** 223–232.

Gale, D., and Sotomayor, M. (1985b). "Ms Machiavelli and the Stable Matching Problem," *Amer. Math. Monthly* **92,** 261–268.

Graham, D. A., and Marshall, R. C. (1987). "Collusive Bidder Behavior at Single Object Second Price and English Auctions," *J. Polit. Econ.*, in press.

Harsanyi, J. C. (1967). "Games with Incomplete Information Played by Bayesian Players. I. The Basic Model," *Manage. Sci.* **14,** 159–182.

Harsanyi, J. C. (1968a). "Games with Incomplete Information Played by 'Bayesian' Players. II. Bayesian Equilibria Points," *Manage. Sci.* **14,** 320–334.

Harsanyi, J. C. (1968b). "Games with Incomplete Information Played by 'Bayesian' Players. III. The Basic Probability Distribution of the Game," *Manage. Sci.* **14,** 486–502.

Kelso, A. S., Jr., and Crawford, V. P. (1982). "Job Matching, Coalition Formation, and Gross Substitutes," *Econometrica* **50,** 1483–1504.

Leonard, H. B. (1983). "Elicitation of Honest Preferences for the Assignment of Individuals to Positions," *J. Polit. Econ.* **91,** 461–479.

McVitie, D. G., and Wilson, L. B. (1970). "Stable Marriage Assignments for Unequal Sets," *BIT* **10,** 295–309.

Mongell, S. J. (1988). *Sorority Rush as a Two-Sided Matching Mechanism: A Game-Theoretic Analysis.* Ph.D. dissertation, Department of Economics, University of Pittsburgh.

Mongell, S. J., and Roth, A. E. (1988), "Sorority Rush as a Two-Sided Matching Mechanism," mimeo.

Myerson, R. B. (1985). "Bayesian Equilibrium and Incentive-Compatibility: An Introduction," *in Social Goals and Social Organizations: Essays in Memory of Elisha Pazner* (L. Hurwicz, D. Schmeidler, and H. Sonnenschein, Eds.). Cambridge: Cambridge Univ. Press.

Roth, A. E. (1982). "The Economics of Matching Stability and Incentives," *Math. Oper. Res.* **7,** 617–628.

Roth, A. E. (1984a). "The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory," *J. Polit. Econ.* **92,** 991–1016.

Roth, A. E. (1984b). "Misrepresentation and Stability in the Marriage Problem," *J. Econ. Theory* **34,** 383–387.

Roth, A. E. (1984c). "Stability and Polarization of Interests in Job Matching," *Econometrica* **52,** 47–57.

Roth, A. E. (1985a). "The College Admissions Problem Is Not Equivalent to the Marriage Problem," *J. Econ. Theory* **36,** 277–288.

Roth, A. E. (1985b). "Common and Conflicting Interests in Two-Sided Matching Markets," *Europ. Econ. Rev.* (Special issue on Market Competition, Conflict, and Collusion) **27,** 75–96.

Roth, A. E. (1986). "On the Allocation of Residents to Rural Hospitals: A General Property of Two-Sided Matching Markets," *Econometrica* **54,** 425–427.

Roth, A. E. (1989). "A Natural Experiment in the Organization of Entry-Level Labor Markets: A Game-Theoretic Analysis of Regional Markets for Physicians in the U.K.," in preparation.

Roth, A. E., and Sotomayor, M. (1988a). "The College Admissions Problem Revisited," *Econometrica,* in press.

Roth, A. E., and Sotomayor, M. (1988b). "The Two-Sided Matching: A Study in Game-Theoretic Modelling and Analysis," mimeo.

Shapley, L. S., and Shubik, M. (1972). "The Assignment Game. I. The Core," *Int. J. Game Theory* **1,** 111–130.

Sotomayor, M. (1987). "The Assignment Game with Multiple Partners." Working paper, Department of Mathematics, Pontificia Universidade Catolica do Rio de Janeiro.